Méthode de décomposition basée sur les statistiques d'ordre

Aleksej Makarov

Traduction française **Loredana Mauriand**

Laboratoire de Micro-informatique
Ecole Polytechnique Fédérale de Lausanne, Suisse
e-mail: aleksej.makarov@di.epfl.ch

RÉSUMÉ

Cet article présente une méthode de décomposition des séries chronologiques en composantes tendancielles et cycliques. La longueur moyenne du cycle des composantes cycliques n'est pas connue a priori. L'approche proposée ici est basée sur un suivi simultané des statistiques d'ordre, à savoir des maxima et des minima locaux, extraits à partir des fenêtres glissantes de différentes longueurs. Contrairement aux méthodes basées sur la transformée de Fourier, la longueur moyenne du cycle peut être estimée, en dépit de la composante tendancielle, permettant ainsi la décomposition du signal. La même procédure définit un test de stationnarité.

ABSTRACT

This paper presents a method for analyzing composite time series. The latter consist of a trend and one or several cyclic components. The average cycle width is not *a priori* known, and cannot be estimated by spectral analysis methods. This precludes the use of traditional decomposition techniques. In the proposed approach, the cycle width is evaluated by tracking a nonlinear combination of rank order statistics, locally estimated from a number of increasing size windows. An iterative scheme based on this estimate is proposed for decomposition of signals. A by-product of the cycle width evaluation procedure is a trend detection method.

1 Introduction

Un signal composite est un signal non-stationnaire composé d'une tendance, et d'une ou plusieurs composantes cycliques. Une composante cyclique est un processus quasi-périodique bruité. Les méthodes classiques (les modèles linéaires dynamiques bayésiens, les modèles SARIMA, et les bancs de filtres aux fréquences de coupure présélectionnées) décomposent ce type de signal en utilisant la longueur du cycle (la période locale) connue a priori. Cependant, ces méthodes échouent lorsque cette valeur n'est pas disponible. Si la tendance sous-jacente n'est pas localement stationnaire (pente arbitraire, différente de zéro), son contenu spectral peut noyer le spectre d'une composante cyclique. Dans ces conditions, les techniques spectrales, telle que la transformée de Fourier à court terme, échouent à fournir l'information sur la longueur du cycle.

Récemment nous avons proposé une nouvelle approche afin de déterminer la longueur moyenne du cycle d'une composante cyclique [2]. Cette approche est basée sur une extraction des extrema locaux (ou d'autres statistiques d'ordre) à plusieurs échelles. Un ajustage très sommaire du seul paramètre de la méthode nous a permis de l'appliquer à la décomposition des signaux composites, tels que les rythmes cardiaques [3] ou les données économiques. Ce paramètre, M_{max}, est une limite supérieure de la longueur du cycle. Pour une composante cyclique strictement périodique, avec une période T, perturbée par un bruit blanc et superposée sur une large gamme des pentes, il peut être démontré que la longueur moyenne du cycle $\bar{T}$, converge vers T avec le nombre d'observations N. La méthode n'est donc pas biaisée. Les changements abruptes de la tendance n'influencent pas l'estimation de T d'une manière significative. Lorsque la pente de la composante tendancielle et l'amplitude de la composante cyclique sont égales à zéro (hypothèse nulle), la probabilité de détecter un signal composite (probabilité de fausse alarme) s'avère indépendante de la distribution du bruit blanc. Par conséquent, la méthode peut être considérée comme non-paramétrique.

2 Description de la méthode

2.1 Points caractéristiques

Un cycle peut être représenté comme une suite des micro-tendances aux pentes non-nulles et nulles. Par exemple, un cycle sinusoïdal contient une pente non-nulle (positive), une pente nulle (maximum), une pente non-nulle (négative) et une pente nulle (minimum). Les cycles, tels qu'on les trouve dans la nature, sont souvent bruités, de manière à ce que leurs micro-tendances intrinsèques soient difficilement répérables. Toutefois, on peut extraire les micro-tendances monotones (pentes non-nulles) et trouver les instants ou elles changent le signe de leur pente (pentes nulles). Ces instants seront désignés comme des *points caractéristiques*.

Les points caractéristiques peuvent être détectés par une fenêtre glissante. Supposons que cette fenêtre contienne M+1 échantillons, couvrant un intervalle de temps Δt. Une micro-tendance ascendante est détectée comme une suite partielle-

ment ordonnée de $M+1$ échantillons :

$$\overbrace{\underbrace{x(k-M)}}^{\min\{x\}_{t-\Delta t}} \quad \underbrace{x(k-M+1) \quad \ldots\ldots\ldots \quad x(N)}_{\text{fenêtre précédente } \{x\}_{t-\Delta t}} \quad \overbrace{x(k)}^{\max\{x\}_t} \qquad (1)$$

Pour la détection d'une micro-tendance descendante, les extrema dans la suite (1) devraient être remplacés par leur contraires respectifs. Une fonction binaire, $i(t)$, peut être définie comme $i(t)=1$, si une micro-tendance ascendante ou descendante est détectée à l'instant t, sinon $i(t)=0$. Une autre fonction binaire, $j(t) = i(t-1) \wedge i(t)$, détecte les instants ou la micro-tendance cesse d'être monotone. Les valeurs du signal où $j(t)=1$ seront appelés les *points caractéristiques* du signal.

2.2 Longueur moyenne du cycle

La figure 1 présente une sinusoïde, $a\sin(2\pi t/T)$, superposée sur une droite, $bt+c$. La période, l'amplitude et la pente doivent satisfaire la contrainte $|b| \leq \frac{2\pi a}{T}$. Supposons que le point D soit l'échantillon du signal à l'instant $t-1$. Le point E est un extremum local du signal. Pour que le point suivant, H, soit détecté comme un point caractéristique ($j(t)=1$), il faut que les échantillons de la fenêtre glissante à l'instant $t-1$ soient arrangés comme dans la suite (1). Les fenêtres AE, BE, CE, et DE satisfont cette condition, mais pas les fenêtres ME, NE ou PE. Toutes les fenêtres YE, dont la taille dépasse celle d'une fenêtre XE (ou X et Y sont des points arbitraires, antérieurs à D) d'un multiple entier de la période T, se comporteront de la même manière que la fenêtre XE. Par exemple, la fenêtre AE est d'une période plus longue que la fenêtre CE, satisfaisant également la condition (1). La fenêtre ME ne remplit pas cette condition (car M n'est pas le minimum de l'intervalle ME), et la fenêtre M_1E non plus.

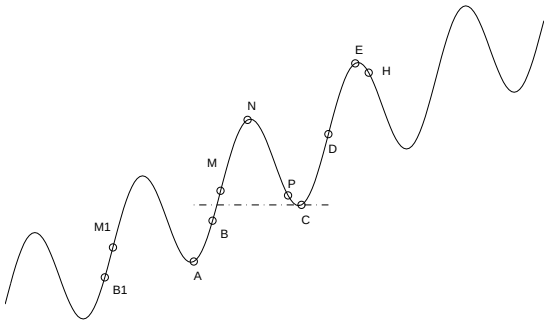

FIG. 1 —

Calculé à partir d'une série de fenêtres dont la taille varie entre $M_{min} \geq 1$ et $M_{max} \gg T$, le nombre de points caractéristiques détectés sur un intervalle de $N > M_{max}$ échantillons présente un comportement cyclique, figure 2. La longueur moyenne du cycle $\bar{T}$ peut être évaluée comme la distance moyenne entre deux sommets avoisinants de la courbe $N(M)$.

Cette méthode, spécialement conçue pour les signaux

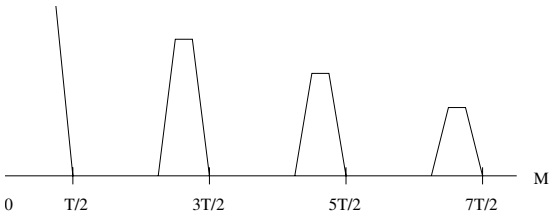

FIG. 2 — *Le nombre de points caractéristiques varie d'une manière cyclique par rapport à la taille M de la fenêtre d'analyse. La longueur moyenne du cycle du signal, $\bar{T}$, peut être mesurée comme la distance entre deux pics avoisinants.*

non-stationnaires, échoue dans l'estimation de la longueur moyenne du cycle dans deux cas :

1. Une périodicité exempte de bruit, oscillant autour d'un niveau constant (b=0). Dans ce cas, les séquences du type (1), ainsi que les points caractéristiques ne peuvent être détectés qu'avec les fenêtres dont la taille μ ne dépasse pas une demi-période, $1 \leq \mu < \frac{T}{2}$. Par conséquent, la courbe $N(M)$ présente un seul pic, et la longueur du cycle $\bar{T}$ ne peut être mesurée comme ci-dessus. Cependant, si le nombre de points caractéristiques par cycle K est connu à l'avance ($K=2$ pour une sinusoïde), la période peut être estimée comme $\bar{T} = KN(\mu)$. Sinon, $2\max_{N(M)>0}(M)$ est aussi une évaluation approximative de la période T. Une évaluation plus précise peut être obtenue par la transformation de Fourier.
2. Une tendance trop raide, $|b| > \frac{2\pi a}{T}$, impliquerait $i(t)=1$ tout au long de sa durée, même là où les points caractéristiques auraient dû être détectés (pour que $j(t)=1$, il faut que $i(t-1)=0$). Heureusement, de telles tendances arrivent rarement et durent brièvement dans la plupart des signaux naturels. Ainsi, la dégradation de la courbe $N(M)$ est insignifiante, et la longueur moyenne du cycle peut être estimée comme la distance moyenne entre deux pics.

2.3 Décomposition

Sitôt trouvée la longueur moyenne du cycle, la composante tendancielle $\tau(t)$ peut être extraite selon une des méthodes classiques de décomposition. A condition d'un faible changement de tendance au cours d'un cycle, cette composante peut être récupérée comme une séquence de valeurs, dont la période d'échantillonnage est égale à la longueur moyenne du cycle,

$$\tau(t_k) = \sum_{i=t_k-\lfloor\frac{\bar{T}}{2}\rfloor}^{t_k+\lfloor\frac{\bar{T}}{2}\rfloor} x(i) \qquad (2)$$

$$t_j - t_{j-1} \approx \bar{T}, \quad i,j,k,t_j, \in \mathcal{N}$$

L'équation (2) suppose que la moyenne de la composante cyclique au cours d'un cycle soit égale à zéro. Afin de suivre les changements de la longueur du cycle, le signal composite peut être réparti en blocs de données de taille N.

Les valeurs de la tendance $\tau(t_k)$ peuvent être estimées pour tous les instants discrets t de l'intervalle observé, par exemple

par l'interpolation linéaire. Dans ce cas, les discontinuités entre deux ségments de taille $\lceil \bar{T} \rceil$ peuvent être "lissées" en appliquant une moyenne mobile (typiquement 3 échantillons de longueur).

Par conséquent, la composante cyclique $\omega(t)$ peut être estimée comme :

$$\omega(t) = x(t) - \tau(t) \tag{3}$$

Lorsque plusieurs composantes cycliques sont présentes dans le signal, le procédé décrit par les équations (2) et (3) peut être appliqué de manière répétitive. L'algorithme ci-dessous est appliqué si deux composantes cycliques, $\omega_1(t)$ et $\omega_2(t)$, aux longueurs du cycle T_1 et T_2 différentes, se trouvent dans un signal composite $x(t)$:

1. Diviser signal $x(t)$ dans des blocs de données de N échantillons, $N \gg \max(T_1, T_2)$.
2. Estimer la longueur moyenne du cycle T_1 de la composante $\omega_1(t)$, $T_1 < T_2$, pour chaque bloc de $x(t)$.
3. Extraire la tendance $\tau_1(t_k)$ selon l'équation (2), et par l'interpolation linéaire suivie de lissage, la tendance $\tau_1(t)$.
4. Extraire $\omega_1(t) = x(t) - \tau_1(t)$.
5. Estimer la longueur moyenne du cycle T_2 pour chaque bloc, à partir de $\tau_1(t)$.
6. Extraire la tendance $\tau(t_k)$,

$$\tau(t_k) = \sum_{i=t_k-\lfloor \frac{\bar{T}_2}{2} \rfloor}^{t_k+\lfloor \frac{\bar{T}_2}{2} \rfloor} \tau_1(i)$$

   et par l'interpolation linéaire suivie de lissage la tendance $\tau(t)$.
7. Extraire $\omega_2(t) = \tau_1(t) - \tau(t)$.

La même démarche est appliquée au cas où plusieurs composantes cycliques sont présentes dans le signal observé.

3 Détection des tendances

Si un signal est composé d'une tendance et du bruit blanc, les pics de $N(M)$ (nombre de points caractéristiques en fonction de la longueur de fenêtre) ne sont pas équidistants. Tout de même, la distance moyenne entre deux pics peut toujours être calculée. Cette valeur ne doit pas être confondue avec la longueur moyenne du cycle d'une composante cyclique. Une composante cyclique peut être distinguée du bruit non-corrélé, en observant la variance des distances entre deux pics avoisinants de $N(M)$.

Un signal dépourvu de composantes tendancielles et cycliques est réduit à un niveau constant, perturbé par le bruit. Les points caractéristiques détectés dans un tel signal ne comportent pas d'information importante. Une analyse du nombre de points caractéristiques permet de déceler les niveaux constants bruités. On considère un niveau constant bruité comme l'hypothèse nulle, et une tendance bruitée comme l'hypothèse alternative.

La probabilité de détection des points caractéristiques dans l'hypothèse nulle peut être déduite de la séquence (1). Comme le bruit est supposé être non-corrélé, la probabilité que le dernier échantillon dans la suite soit un maximum est $\frac{1}{M+1}$. Si c'est le cas, la probabilité que le premier échantillon dans la suite soit le minimum est $\frac{1}{M}$. Donc, une micro-tendance (à pente non-nulle) serait détectée avec la probabilité suivante :

$$P_T = \frac{2}{(M+1)M} \tag{4}$$

Par conséquent, un point caractéristique est détecté avec la probabilité

$$P_c = P_T(1 - P_T), \tag{5}$$

indépendamment de la distribution du bruit.

Supposons que le nombre de points caractéristiques soit N. Pour chaque fenêtre, la détection des points caractéristiques ne peut commencer qu'à partir de l'échantillon M+2. Le nombre de points caractéristiques détecté dans le bruit blanc est donc :

$$N_F = P_c(N - M - 1) \tag{6}$$

La figure 3 présente la courbe $N_F(M)$. L'écart entre cette courbe théorique et le nombre actuel de points caractéristiques, $N(M)$, est mesuré par le test χ^2. Ainsi, l'hypothèse nulle peut être rejetée ou acceptée avec une certaine probabilité.

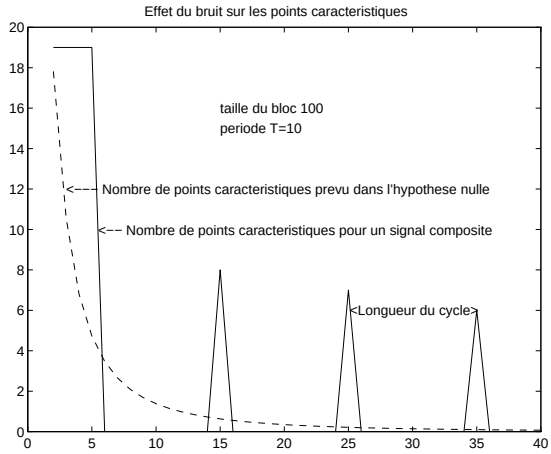

FIG. 3 — *L'évolution du nombre de points caractéristiques, par rapport aux différentes tailles M de la fenêtre glissante, permet de détecter les signaux dépourvus de composantes tendancielles et cycliques.*

4 Illustration

Les données économiques présentent souvent des variations saisonnières ou cycliques. Ainsi, les ventes de voitures augmentent généralement au printemps. Le premier diagramme de la figure 4) présente les ventes mensuelles de voitures aux Etats-Unis entre 1978 et 1984. Le périodogramme classique, basé sur la transformée de Fourier, ne permet pas de mettre en évidence ces variations annuelles. En effet, le spectre

de la composante tendancielle peut cacher les pics discrets correspondant aux cyclicités. Ainsi, dans le périodogramme (second diagramme de la figure 4), le pic le plus élevé correspond aux cycles trisannuels. Ce pic résulte des variations de la tendance. Le deuxième pic est également observé pour ce type de signaux, afin de réduire l'influence de la tendance. Dans notre cas, celui-ci indique une fréquence qui avoisine les cycles semestriels, donc une harmonique des variations saisonnières. Pour l'enregistrement de 1978 à 1988, le deuxième pic aurait permis de détecter les cycles annuels. Dans le périodogramme de la série chronologique analysée (1978-1984), les cycles annuels ne sont pas distincts.

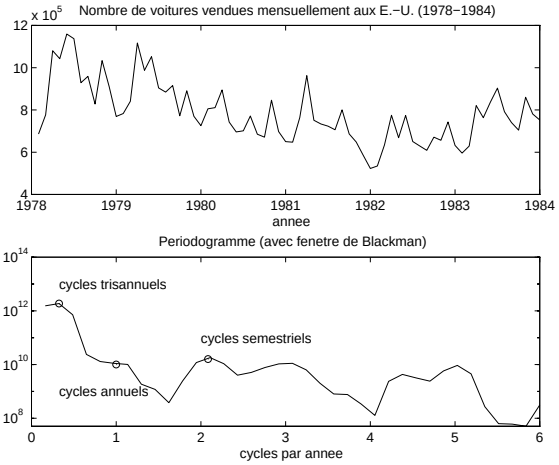

FIG. 4 — *Le périodogramme de Fourier échoue à estimer la fréquence des cycles dominants superposés sur une tendance.*

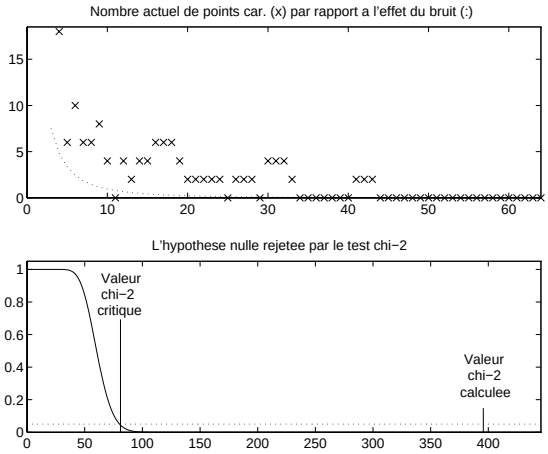

FIG. 5 — *Le nombre de points caractéristiques $N(M)$ (en haut) peut être utilisé comme un test de détection des tendances. Une comparaison (le test χ^2) entre le $N(M)$ actuel et celui prévu par la théorie dans l'hypothèse nulle, permet d'établir le caractère composite du signal.*

En balayant la même série chronologique (en un seul bloc) par une série de fenêtres de trois à soixante-quatre échantillons, on découvre qu'il s'agit d'un signal composite, l'hypothèse nulle étant rejetée, figure 5. La distance moyenne entre les pics avoisinants (premier diagramme de la figure 5

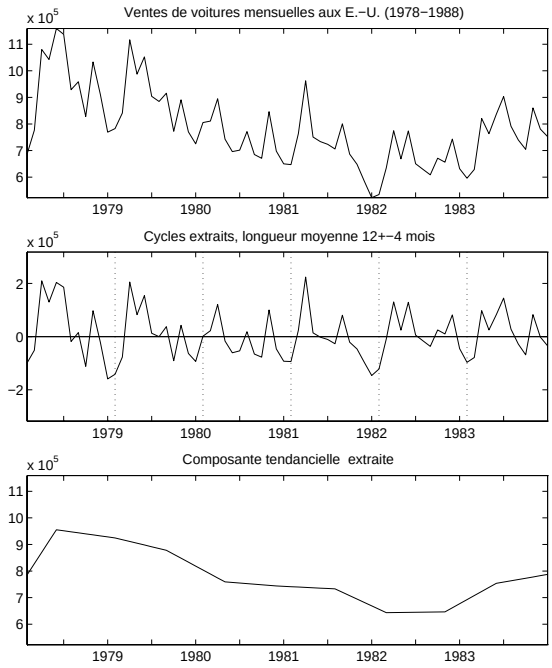

FIG. 6 — *Le signal original (en haut), les cycles extraits (au milieu) et la tendance extraite (en bas).*

est de douze mois, ou plus précisement 12 ± 4 mois, dans les intervalles de confiance de 95.4%. Donc, les cycles annuels ont pu être mis en évidence. Pour l'enregistrement de 1978 à 1988, l'écart-type de la longueur moyenne du cycle passe de deux mois à un mois et demi. Le signal est décomposé en utilisant la longueur moyenne du cycle, figure 6.

Références

[1] M.West. Bayesian inference in cyclical component dynamic linear models. *J. Am. Statist. Ass.*, 90(432) :1301–1312, December 1995.

[2] A. Makarov. Periodicity retrieval from nonstationary signals. *Proc. EUSIPCO*, pp. 1949–1952, 1996.

[3] A. Makarov et G. Thonet. Rank order based decomposition and classification of heart rate signals. *Proc. ECSAP*, Prague, 1997.

[4] T.E.Dielman. *Applied Regression Analysis for Business and Economics*. PWS-KENT, Boston, 1991.

[5] C.W.J.Granger et R.Engle. Applications of spectral analysis in economics. *Time Series in the Frequency Domain*, pp. 93–110. North-Holland, 1983.